Capitolo 5

Valutazione dell'errore numerico

5.1 Errore numerico

Il calcolo di una derivata o la soluzione di un problema differenziale attraverso le metodologie tipiche della analisi numerica introduce un errore detto errore numerico che dipende dal passo di discretizzazione. In particolare, il valore esatto della derivata o della soluzione g del problema differenziale differisce da quello calcolato numericamente $\overline{g}$ per l'errore numerico ε_n.

$$\overline{g} - g = \varepsilon_n \tag{5.1}$$

Questo errore può essere ridotto, riducendo il passo di discretizzazione del metodo numerico utilizzato, fino a raggiungere il livello di errore desiderato.

In generale, quando si risolve un problema differenziale o si calcola una derivata con un metodo numerico non si conosce la soluzione esatta e, pertanto, per quantificare l'errore numerico è necessario calcolare una stima della soluzione esatta. Esiste un procedimento che permette di calcolare una stima della soluzione esatta utilizzando le soluzioni numeriche ottenute dalla stessa formula alle differenza o dallo stesso schema di integrazione con passi di discretizzazione differenti. Questo procedimento prende il nome di estrapolazione alla Richardson.

5.2 Errore numerico di una funzione $\overline{g}(x)$

Se si considera la derivata di una funzione della sola variabile indipendente (per esempio x) o un problema differenziale alle derivate ordinarie, allora il valore $\overline{g}$ ottenuto attraverso una formula alle differenze o attraverso uno schema di integrazione numerica in un punto fissato x differirà da quello esatto g per

l'errore ε_n che, in generale si può esprimere come

$$\varepsilon_n = \sum_{i=n}^{\infty} k_i (\Delta x)^i \tag{5.2}$$

dove n è l'ordine formale della formula alle differenze o dello schema numerico utilizzato, k_i il coefficiente dell'i-esimo termine dell'errore, Δx il passo di discretizzazione utilizzato dal metodo numerico. Ciascun termine dell'errore numerico tende a zero per $\Delta x \to 0$, tuttavia i termini con $i > n$ tendono a zero più rapidamente del termine con $i = n$. Pertanto, per valori di Δx sufficientemente piccoli si avrà che

$$|k_n (\Delta x)^n| >> \left| \sum_{i=n+1}^{\infty} k_i (\Delta x)^i \right| \tag{5.3}$$

ovvero il primo termine dell'errore fornirà un contributo predominante rispetto a quello fornito dalla somma di tutti gli altri termini. Quando ciò si verifica, allora la convergenza $\overline{g} \to g$ è detta asintotica ed l'intero errore può essere approssimato dal primo termine di errore.

$$\varepsilon_n = \sum_{i=n}^{\infty} k_i (\Delta x)^i \approx k_n (\Delta x)^n \tag{5.4}$$

Per questa ragione questo termine prende il nome di termine dominante dell'errore numerico. E' importante sottolineare che, se Δx è sufficiente piccolo da assicurare la convergenza asintotica, ciò non significa che l'errore sia in assoluto piccolo. Si possono verificare situazioni per le quali l'errore numerico raggiunge valori molto piccoli ma la soluzione numerica non é ancora in convergenza asintotica o, viceversa, situazioni dove la convergenza asintotica si raggiunge quando l'errore numerico é ancora molto grande.

Per ragioni che saranno chiare tra breve, quando si calcola numericamente una derivata o un problema differenziale è auspicabile utilizzare un Δx sufficientemente piccolo tale da assicurare la convergenza asintotica. Tuttavia, in generale, non e' possibile conoscere *a priori* per quale scelta di Δx si verifica la convergenza asintotica. Inoltre, per una determinata formula alle differenze o per un determinato schema di integrazione numerico i valori di Δx a partire dai quali si manifesta la convergenza asintotica dipendono dalla funzione derivata o problema differenziale considerato. Solo dall'analisi delle soluzioni numeriche calcolate con differenti passi di discretizzazione, e quindi *a posteriori*, è possibile verificare se il termine dominante dell'errore numerico è appunto dominate e la convergenza del valore numerico al valore esatto è di tipo asintotico.

5.3 Estrapolazione alla Richardson per una funzione $\overline{g}(x)$

Si calcoli numericamente la derivata o la funzione soluzione del problema differenziale nel punto specificato con una formula o con un schema numerico di ordine n utilizzando tre valori differenti del passo di discretizzazione: Δx, $\Delta x/2$ e $\Delta x/4$. In questo modo si otterranno tre distinte valutazioni del valore g ciascuna relativa ad un livello di discretizzazione differente. In particolare, il valore ottenuto con Δx (il livello di discretizzazione più rado) sarà indicato con $\overline{g}^r$, mentre con $\overline{g}^m$ e $\overline{g}^f$ saranno indicati quelli ottenuti rispettivamente con $\Delta x/2$ e $\Delta x/4$ (i livelli di discretizzazione piú fini). Utilizzando le (5.1) e (5.2), per ciascun valore numerico ottenuto si può scrivere una relazione che lega $\overline{g}$ e g a ε_n

$$\begin{aligned}
\overline{g}^r &= g + K_n (\Delta x)^n + \sum_{i=n+1}^{\infty} K_i (\Delta x)^i \\
\overline{g}^m &= g + K_n \left(\frac{\Delta x}{2}\right)^n + \sum_{i=n+1}^{\infty} K_i \left(\frac{\Delta x}{2}\right)^i \\
\overline{g}^f &= g + K_n \left(\frac{\Delta x}{4}\right)^n + \sum_{i=n+1}^{\infty} K_i \left(\frac{\Delta x}{4}\right)^i
\end{aligned} \tag{5.5}$$

Se si suppone che Δx sia sufficientemente piccolo da garantire la convergenza asintotica, allora nelle (5.5) il termine dominate dell'errore sarà predominante rispetto alla parte rimanente e pertanto l'errore che si commetterà nel trascurare questa parte rimanente dell'errore sarà piccolo. Riscrivendo le tre relazioni (5.5) con solo il termine dominante si ottiene un sistema di tre equazioni nelle tre incognite $\tilde{g}$, $\tilde{K}_n$ e $\tilde{n}$.

$$\begin{aligned}
\overline{g}^r &= \tilde{g} + \tilde{K}_n (\Delta x)^{\tilde{n}} \\
\overline{g}^m &= \tilde{g} + \tilde{K}_n \left(\frac{\Delta x}{2}\right)^{\tilde{n}} \\
\overline{g}^f &= \tilde{g} + \tilde{K}_n \left(\frac{\Delta x}{4}\right)^{\tilde{n}}
\end{aligned} \tag{5.6}$$

Le tre quantità $\tilde{g}$, $\tilde{K}_n$ e $\tilde{n}$ sono state indicate in modo differente dalle analoghe quantità g, K_n e n che compaiono nelle (5.5), poiché nel passaggio dalle relazioni (5.5) alle relazioni (5.6) è stata introdotta una approssimazione. Le quantità $\tilde{g}$, $\tilde{K}_n$ e $\tilde{n}$ tenderanno alle g, K_n e n per $\Delta x \to 0$ ma non potranno (eccetto in qualche caso particolare) essere uguali per valori di $\Delta x \neq 0$.

La soluzione del sistema (5.6) si può calcolare attraverso le seguenti formule:

$$\tilde{n} = \frac{\log \frac{\overline{g}^r - \overline{g}^m}{\overline{g}^m - \overline{g}^f}}{\log 2}$$

$$\begin{aligned} \tilde{K}_n &= \frac{(2)^{\tilde{n}}(\overline{g}_n^m - \overline{g}^r)}{(\Delta x)^{\tilde{n}}\left[(2)^{\tilde{n}} - 1\right]} \\ \tilde{g} &= \overline{g}^r + \tilde{K}_n (\Delta x)^{\tilde{n}} \end{aligned} \tag{5.7}$$

Poiché l'ordine di accuratezza formale n della formula o dello schema di integrazione è noto, è possibile verificare se le tre soluzioni numeriche $\overline{g}^r$, $\overline{g}^m$ e $\overline{g}^f$ sono state ottenute in condizioni di convergenza asintotica confrontando l'ordine formale n con il valore $\tilde{n}$ ottenuto dalle (5.6). Se $\tilde{n}$ è molto prossimo a n allora nel passaggio dalle (5.5) alle (5.6) è stato commesso un piccolo errore e, quindi, la convergenza delle soluzioni numeriche $\overline{g}^r$, $\overline{g}^m$ e $\overline{g}^f$ alla soluzione esatta g è di tipo asintotico. Ma se le tre valutazioni numeriche sono state ottenute in condizioni di convergenza asintotica anche la quantità $\tilde{g}$ sarà molto prossima al valore esatto g (che in generale non è noto). Pertanto, in questo caso, $\tilde{g}$ è una buona stima di g che può essere utilizzata per la stima dell'errore numerico attraverso la relazione

$$\varepsilon_n = \overline{g} - g \approx \overline{g} - \tilde{g} \tag{5.8}$$

5.4 Estensione al caso di una funzione $g(x,t)$

La soluzione numerica $\overline{g}$ di uno dei tre problemi differenziali considerati nei capitoli precedenti ottenuta da uno schema numerico accurato all'ordine j nel tempo e all'ordine m nello spazio è legata alla soluzione esatta g attraverso la relazione

$$\overline{g} = g + \underbrace{K_{tj}(\Delta t)^j + \sum_{i=j+1}^{\infty} K_{ti}(\Delta t)^i}_{\textit{Errore nel tempo}} + \underbrace{K_{xm}(\Delta x)^m + \sum_{i=m+1}^{\infty} K_{xi}(\Delta x)^i}_{\textit{Errore nello spazio}} \tag{5.9}$$

nella quale si può osservare la presenza di termini di errore legati alla discretizzazione nel tempo e nello spazio. Nel caso di convergenza asintotica, si deve avere che

$$\begin{aligned} \left|K_{tj}(\Delta t)^j\right| &>> \left|\sum_{i=j+1}^{\infty} K_{ti}(\Delta t)^i\right| \\ |K_{xm}(\Delta x)^m| &>> \left|\sum_{i=m+1}^{\infty} K_{xi}(\Delta x)^i\right| \end{aligned} \tag{5.10}$$

Anche in questo caso è possibile procedere all'estrapolazione alla Richardson. Tuttavia esistono tre modi differenti di condurre l'estrapolazione.

Nel primo tipo di estrapolazione si utilizzano tre valutazioni differenti di $\overline{g}$ con passo di discretizzazione nello spazio decrescente (Δx, $\Delta x/2$, $\Delta x/4$) e passo di discretizzazione nel tempo (Δt) inalterato.

$$\begin{aligned} \overline{g}^r &= g + K_{tj}(\Delta t)^j + K_{xm}(\Delta x)^m + .. \\ \overline{g}^m &= g + K_{tj}(\Delta t)^j + K_{xm}(\frac{\Delta x}{2})^m + .. \\ \overline{g}^f &= g + K_{tj}(\Delta t)^j + K_{xm}(\frac{\Delta x}{4})^m + .. \end{aligned} \tag{5.11}$$

Se nella (5.11) si trascurano i termini di errore di ordine superiore e si effettuano le seguenti sostituzioni

$$\begin{aligned} \tilde{g} &= g + K_{tj}(\Delta t)^j \\ \tilde{K} &= K_{xm} \\ \tilde{m} &= m \end{aligned} \tag{5.12}$$

si ottiene un sistema identico al (5.6), la cui soluzione è data dalle (5.7). Tuttavia in questo caso $\tilde{g}$ non è una stima del valore esatto g, poichè $\tilde{g}$ contiene anche la parte di errore relativa alla discretizzazione nel tempo. Questo tipo di estrapolazione permette di stimare solo la parte dell'errore legata alla discretizzazione nello spazio, ma non può fornire una stima dell'errore numerico globale. Inoltre, questa analisi verifica l'ordine di accuratezza nel tempo e determina a partire da quali valori di Δx si verifica la convergenza asintotica per la parte di errore legato alla discretizzazione nello spazio.

Il secondo caso è analogo al primo caso, in quanto si utilizzano tre valutazioni di $\overline{g}$ ottenute riducendo il passo di discretizzazione nel tempo (Δt, $\Delta t/2$ e $\Delta t/4$) e lasciando inalterato il passo di discretizzazione nello spazio (Δx). In questo caso si ottiene un sistema simile al (5.6)

$$\begin{aligned} \overline{g}^r &= \tilde{g} + \tilde{K}(\Delta t)^{\tilde{j}} \\ \overline{g}^m &= \tilde{g} + \tilde{K}(\frac{\Delta t}{2})^{\tilde{j}} \\ \overline{g}^f &= \tilde{g} + \tilde{K}(\frac{\Delta t}{4})^{\tilde{j}} \end{aligned} \tag{5.13}$$

dove $\tilde{g}$ e' una stima della somma della soluzione esatta più l'errore relativo alla discretizzazione spaziale, mentre $\tilde{j}$ è una stima del ordine di accuratezza nel tempo dello schema numerico.

Il terzo tipo di estrapolazione è quello più interessante poiché permette di valutare la stima dell'errore numerico globale. In questo caso si considerano valutazioni numeriche ottenute riducendo sia il passo di discretizzazione nel tempo che quello nello spazio in modo da rispettare la seguente relazione

$$\frac{(\Delta t)^j}{(\Delta x)^m} = A \tag{5.14}$$

essendo A una costante. In particolare se si considera seguenti i seguenti passi di discretizzazione spaziale

$$\Delta x; \quad \frac{\Delta x}{2}; \quad \frac{\Delta x}{4}$$

ad essi dovranno essere associati i seguenti passi di discretizzazione temporale

$$\Delta t; \quad \frac{\Delta t}{2^{m/j}}; \quad \frac{\Delta t}{4^{m/j}}$$

Poichè tra Δx e Δt esiste la relazione (5.14) allora la (5.9) si può riscrivere come

$$\overline{g} = g + K_{tj}(A)^{1/j}(\Delta x)^m + K_{xm}(\Delta x)^m + .. = g + K_m(\Delta x)^m + .. \qquad (5.15)$$

avendo posto $K_m = K_{tj}(A)^{1/j} + K_{xm}$. Pertanto dalle tre valutazioni numeriche ottenute con i seguenti passi di discretizzazione

$$\begin{aligned} (\Delta x, \Delta t) &\rightarrow \overline{g}^r \\ \left(\frac{\Delta x}{2}, \frac{\Delta t}{2^{m/j}}\right) &\rightarrow \overline{g}^m \\ \left(\frac{\Delta x}{4}, \frac{\Delta t}{4^{m/j}}\right) &\rightarrow \overline{g}^f \end{aligned}$$

si ottiene il seguente sistema

$$\begin{aligned} \overline{g}^r &= \tilde{g} + \tilde{K}_m(\Delta x)^{\tilde{m}} \\ \overline{g}^m &= \tilde{g} + \tilde{K}_m(\frac{\Delta x}{2})^{\tilde{m}} \\ \overline{g}^f &= \tilde{g} + \tilde{K}_m(\frac{\Delta x}{4})^{\tilde{m}} \end{aligned} \qquad (5.16)$$

dove in questo caso $\tilde{g}$ è una stima della soluzione esatta che può essere utilizzata per determinare la stima dell'errore numerico globale.